Dynamic Sign Language Recognition based on Depth Motion Volumes and Key Frames

Yong HU
School of Software Engineering, Jinling Institute of Technology, Nanjing, 211169, Jiangsu, China
E-mail: huyong@jit.edu.cn

Article Info
Volume 83
Page Number: 574 - 585
Publication Issue:
July-August 2020

Article History
Article Received: *06 June 2020*
Revised: *29 June 2020*
Accepted: *14 July 2020*
Publication: *25 July 2020*

Abstract

This paper proposes a dynamic sign language recognition approach through the use of Depth Motion Volumes (DMV). Different from the previous research by stacking the difference of the motion energy of depth images, a novel strategy is applied for forming DMV. The selected key frames of the entire video sequences are stacked in chronological order and then projected into three orthogonal planes. Distinctive features are calculated by Histograms of Oriented Gradients (HOG) and then input to LIBSVM for recognition. The experiments are executed on Microsoft Research Gesture3D dataset and the results indicate that the proposed approach is efficient and performs better than the compared approaches

Keywords: *Dynamic Sign Language Recognition, Depth Motion Volumes, Key Frames, Histograms of Oriented Gradients, Support Vector Machine*

1. Introduction

Sign language (SL) is the major communication approach using between the deaf and dumb people. It is completely natural and intuitive language, uses hand gestures and facial expressions to express meanings. The objective of automatic sign language recognition (SLR) are: remove the huge communication obstacle between deaf and hearing people, and bring a freely and naturally experience of human-computer interaction. In recent years, SLR based on computer vision has become a cross-disciplinary research hotspot in the new human-computer interaction. The research of sign language recognition began in the early 1990s. In the past three decades, many sign language recognition approaches have been proposed to improve the recognition rate and speed. Lots of existent surveys were presented to show the improvements in this area [1-3].

Initially, sign language recognition mainly uses the direct detection of machine equipment to obtain the spatial information of human hands and joints, such as hand's position, angle, and the location of the fingertips. This kind of approaches are called sensor-based approach and the typical representative devices are data gloves. But data gloves are limited by the naturalness, complicated and expensive, and were soon replaced by optical marking. The optical marking approaches can also provide good results, but it still needs more complex devices. Compared with wearable device recognition system, vision-based sign language recognition system can enable signers to interact with each other freely and naturally. Because of its characteristic, the widely used visual sign language recognition has become a hot research topic in computer vision at present and in the future.

In SL system, hand(s) movement, body postures and face expression are used to express linguistic meanings. In generally speaking, sign language gestures includes two parts, static gesture and dynamic gesture. Static gestures are represented by an individual hand or finger shape, while dynamic gestures are represented in a sequence of hand movement trajectories. The difficulties mentioned below all make the task of sign language recognition very challenging. The major challenges of SLR lie in:

(1) The visual similarity in global trajectory of some signs.

(2) The invisibility of the finger(s) that occurs in signing.

(3) The large vocabulary of signs in the sign language dictionary.

(4) The large amount of variation by different signers (signer dependent variations).

(5) The complex background environment.

In the last few years, depth camera has been more and more popular in the field of sign language recognition. Based on structured light illumination technology, depth cameras provide an opportunity for sign language recognition based on depth images. As one of the favorite somatosensory peripheral, Microsoft Kinect ™ was extensively applied as visual input device. Based on structured light illumination technology, colour and depth information at high resolution can be obtained by Kinect sensors. The multiple channel information are proved to achieve higher recognition rate. Due to the advantage of insensitive to the change of light illumination, more and more research on action & gesture recognition system are implemented by using depth information.

Zhang and Tian et al [4] proposed a novel and effective descriptor, the Histogram of 3D Facets (H3DF), to encode the 3D shape information from depth maps. And then combining the H3DF with an N-gram model and dynamic programming. The proposed descriptor is extensively evaluated on two public 3D static hand gesture datasets, one dynamic hand gesture dataset, and one popular 3D action recognition dataset. Yang et al [5] proposed a fast algorithm for computing the likelihood of Hidden Markov model (HMM) to calculate the similarity between the sign model and testing sequence. Experiments were conducted on a Kinect dataset of Chinese sign language containing 100 sentences composed of 5 signs each. Plouffe and Cretu [6] adopted a block search scheme and k-curvature algorithm to locate the hand position, and recognize gestures by using the Dynamic Time Warping (DTW) algorithm. Experiments were obtained on nine static gestures, the ASL (American Sign Language) alphabet and a few dynamic gestures. Lee [7] proposed a Kinect-based recognition system. Hand positions, hand signing direction, and hand shapes are extracted from the signing gestures. A Support Vector Machine (SVM) classifier is trained for Taiwanese sign language recognition. Kumar et al [8] proposed a novel multi-sensor fusion framework using Coupled Hidden Markov Model (CHMM) for sign language recognition. The experiments dataset consists of 25 dynamic sign words of Indian Sign Language. Huang et al [9] proposed a novel sequence-to-sequence learning method based on Keyframe Centred Clips (KCCs). The empirical results were obtained on a dataset containing 310 Chinese sign language words.

The presented work concentrates on recognizing dynamic sign language. A new representation of dynamic signs, Depth Motion Volumes, is proposed for the perspective of computational efficiency. The selected key frames of the entire video sequences are stacked in chronological order and then projected into three orthogonal planes. Distinctive features are calculated by project DMV into three orthogonal planes (*x*-*y*, *y*-*z* and *z*-*x*). Combining with the features extracted by HOG descriptor from key frames, the recognition stage is performed by LIBSVM in this work. The rest of the paper is organized as follows. In Section 2, the works of related literature are presented and discussed. The details of generating DMV and key frame selection are presented in Section-3. The preprocessing step, feature extraction and LIBSVM setup are presented in Section-4. Results are presented in Section 5. Finally, the conclusion and future work are presented in Section 6.

2. Related Literature

2.1 Depth Motion Maps

In the last few years, the emergence of RGB-D cameras provides an opportunity for action & gesture recognition based on depth information. Based on structured light illumination technology, 3D information can be captured for recognition. Due to the advantage of insensitive to the change of light illumination, more and more research on action & gesture recognition system are implemented by using depth information.

As one of the spatio-temporal features, Depth Motion Maps (DMMs) are widely used as descriptive features for human action recognition. The DMMs can be formed in two steps. Firstly, pile up the motion energy of foreground regions of depth images, the stacked motion maps indicate the movement and variation of the actions. Secondly, cast the motion maps onto 3 rectangular planes to generate a DMMs. The particular visual aspect of each DMMs can be used as descriptive features for recognition. DMM from a video sequence with N frames is obtained as follow equation.

$$DMM = \sum_{n=2}^{N} |img_p^n - img_p^{n-1}| \qquad (1)$$

where: n is the frame number; img_p^n is the map of nth frame image from projection view - front, side, or top.

Yang et al [10] proposed an efficient and effective method for human action recognition based on DMMs. The depth images in video sequence were casted onto three projective views (front, top and side view) to generate a DMMs. The divergence between adjacent images were computed and thresholded, and then as the motion energy of every projected map. Histograms of Oriented Gradients (HOG) are computed from DMMs as descriptive features. The recognition results on Microsoft Research (MSR) Action3D dataset show significantly outperforms of the approach.

Chen et al [11] proposed a DMMs based approach for human action recognition. The difference between the proposed approach and [10] is the computational way of the motion energy. The absolute divergence between adjacent images were computed without thresholding and accumulated to form a DMMs. The recognition results on MSR Action3D dataset show superior performance of the proposed approach.

2.2 Key Frames

Video sequences for dynamic sign language recognition systems usually last several seconds and thus contains dozens consecutive frames. In traditional recognition systems, all the frames of the sequences are processed and applied for recognition. But recent researches show the fact that only few essential and distinguishing frames called key frames are representative and other frames are unnecessary for processing. It will cause more ambiguities and low efficiency if considering all the frames of video sequences. Therefore, the key frames extraction method is widely used in dynamic action and gesture recognition systems. The advantages of using key frames lie in: not only speed up the computation, but also reduce useless information.

Tripathi et al [12] proposed a key frame extraction approach from video sequences by measuring transformation in hand shape or position of the signs. Gradient-based algorithm was adopted for selecting key frames from depth video sequences. The frame was first segmented and then image gradient was calculated. Adjacent frames with constant gradient represent the end or the beginning of gestures. Rokade et al [13] identify key frames as the middle frames of two consecutive frames of the maximum hand movement. The concept is based on the fact that maximum hand movement only occurs at the end of one sign and the beginning of another sign. Consider a video sequence with T frames, the movement (difference) was calculated as follows.

$$\text{diff}_i = \sum_{i=1}^{T-1} (frame_{i+1} - frame_i) \qquad (2)$$

where: i is the frame number; $frame_i$ is the pixel-wise image of the video sequence.

Zare and Zahiri [14] consider key frames are generally occurred in small variation states. Consider a video sequence with T frames, the subtractions Sub_i between RBG frames were calculated as follows.

$$Sub_i = \sum_{i=1}^{T-1} |img_{i+1} - img_i| \qquad (3)$$

where: i is the frame number; img_i is the pixel-wise image of the video sequence. The minimum subtraction among several subtractions is first picked out and then compared with a pre-defined threshold. The frame is considered as the key frame only if the subtraction is smaller than the threshold.

3. Proposed Feature Extraction Method

The proposed recognition approach is composed of three stages. Figure 1 shows the main framework. Firstly, depth image of each frames are converted to binary images and the foreground sign regions are separated. And then, a key frame selection method is adopted to reduce the computational complexity. The segmented sign regions of the selected frames are stacked in chronological order to form a DMV. Discriminative features are extracted by using histogram of oriented gradient. In the last stage, LIBSVM is applied for recognition and the experiment results on MSR Gesture3D dataset are compared to the latest approaches.

The computation details of the proposed feature representation, Depth Motion Volumes, is demonstrated in this section. Different from the previous research by stacking the difference of the motion energy of depth images, a novel strategy is applied for forming DMV. The unique shape patterns and properties of the 3D object surfaces expose plentiful distinctive information for characterizing category. The compact representation extracted from DMV proves the proposed approach is sufficient to achieve accurate results.

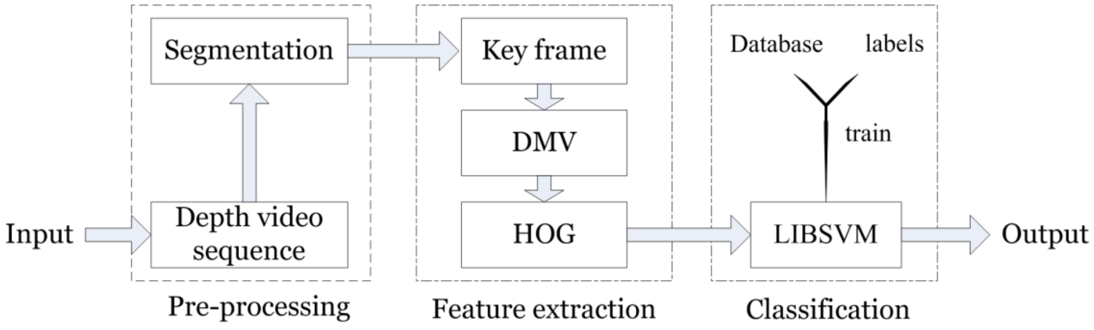
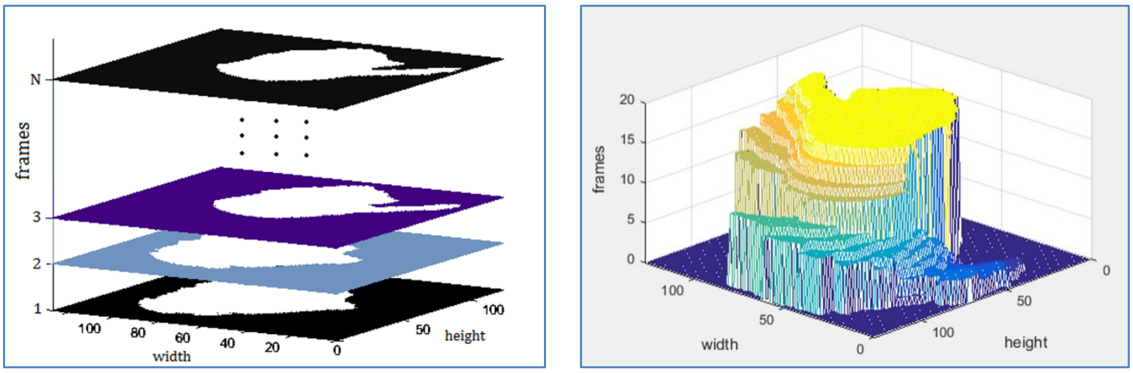


Figure 1. The framework of the proposed approach

3.1 Depth Motion Volumes based features

In order to indicate where the sign motions are occurred, a novel strategy is applied for forming DMV. For a given depth video sequences, depth image of each frames are first converted to binary images. The foreground sign regions are segmented from each frames. The pixels of sign region are marked as 1 and the background pixels are marked as 0. Then the segmented sign regions of the entire video sequences are stacked in chronological order to form a DMV. Figure 1 shows the DMV generated from video sequence of ASL *J*.

Figure 2(a) shows the diagram of how the frames of video sequence are stacked. In a three orthogonal Cartesian planes, frames are piled up along Z-axis in frame sequence. Each layer is occupied by one binary frame image. In order to emphasize the gradation of each layer, background of some images are fulfilled in other colors. Figure 2(b) shows the practical results of the stacked frames. Owing to the fact that sign region in each layer has the gray value of 1, a three-dimensional polyhedron called DMV is formed by the multiple layer of images. It is clearly observed that DMV has some stair-stepping variations, which indicate the movement variation of the sign sequences. The unique shape properties of the DMV surface reveal abundant distinctive information for recognition. It should be noticed that only key frames of the image sequence are applied for forming a DMV. The key frame selection method will be analyzed in the next subsection.



(a) (b)

Figure 2. The DMV generated from video sequence of ASL *J*

The features extracted from DMV are calculated as follow steps. The DMV is first projected onto three planes (*x-y, y-z, z-x*) by counting the non-zero pixels of all frames. HOG descriptors are then calculated from the projected images. Finally, three feature vectors are concatenated as the final representation of the sign sequences.

3.2 Key Frames Selection

To speeds up the recognition system, a key frame selection method is adopted in this paper. The selection method is based on the facts that

1.When a meaningful sign is played, there will be a slight pause of the hand(s). In this time interval, the relatively stable states will lead to some familiar frames with little difference. So one of the frames can be selected as key frame, and other frames can be considered as redundant frames.

2.The changing process of the dynamic signs will lead to the variations in position movement or shape appearance. If the signs varies too much between consecutive frames, the frames can be considered as key frames.

In this work, a key frame selection method is proposed for extracting the representative frames. Given a sign video sequences, the difference between consecutive frames are first calculated through all frames. The strategy for calculating difference D_n is represented in Equation 4.

$$D_n = \sum_{i=1}^{M}\sum_{j=1}^{N}(frame^n \oplus frame^{n+1}) \quad (4)$$

where n represents the frame index; the symbol $\oplus$ represents the XOR operation; M and N represent the height and the width of the frame image. The difference value is then compared to a presetting threshold. Considering the abovementioned facts, the frame n and $n+1$ are selected as key frames if the difference is greater than the threshold. The number of frames is significantly decreased after key frame selection method, thus the computational complexity of the system is greatly reduced.

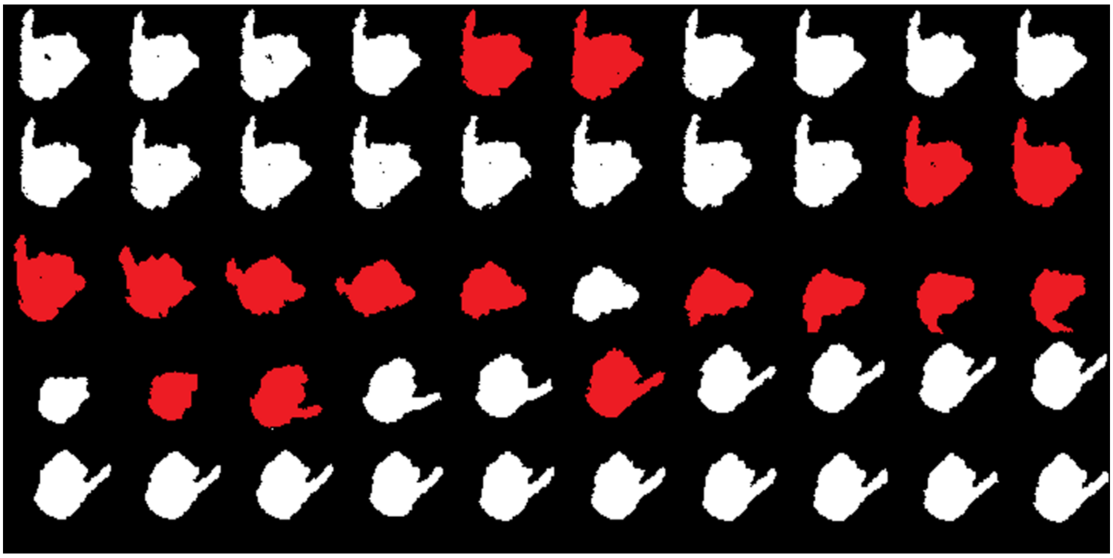


Figure 3. The key frame selection result of ASL *J*

For most of the signs, about 1/3 to 3/5 frames are selected as key frames. The result of key frame selection is based on the calculation strategy and threshold. Figure 3 shows the frames of video sequence of ASL *J,* selected key frames are about 1/3 of total frames and marked with red color. It is observed that there are many familiar frames at the beginning and the ending of the sequence. The familiar frames with little difference are redundant, only cause more computational cost. By using the proposed key frame selection method, the representative key frames greatly reduced the computational complexity of the system.

4. Experimental Results and Discussion

4.1 Dataset

To validate the effectivity of the proposed approach, a benchmark dataset for depth-based gestures recognition system is adopted in this work. The MSR Gesture3D dataset was collected by Wang [15] and applied in many researches [4, 16-17]. The dataset was captured by Kinect sensor, and consists of twelve dynamic ASL gesture signs such as "*blue*", "*milk*", "*where*", "*J*", etc. Each gesture sign is executed two or three times by ten signers, formed a total of 336 depth video sequences. For all the frames, the hand region has been segmented previously. It should be noted that the video sequences of sign "*Pig*" performed by the eighth signer are all blank. Thus there are total 333 video sequences applied for recognition.

4.2 Experimental Results

To evaluate the proposed approach reasonably, the experiments are conducted and compared to the state-of-the-art approaches. Two experiments are conducted in this work, the first one is by using DMV only and the second is by using DMV plus key frames (see Table 1). Radial Basis Function (RBF) are utilized in LIBSVM [18] by using ten times cross validations. Followed the same settings in [4, 16-17], the leave-one-out strategy is adopted in this work. For each experiment, recordings of 9 signers are applied for training and all recordings for testing. Results list in Table 1 is the average accuracy of the experiments. It can be seen from the table, the proposed DMV performs slightly better than Zhang (2015) and Chen (2018). The performance by using DMV plus key frames achieved a higher recognition rate at 96.1%. Experiments results indicate that the proposed approach is efficient and performs better than the compared approaches.

Table 1. Performance of difference method

method	**average accuracy (%)**
[4] Histogram of 3D Facets	95.0
[15] action graph	88.5
[16] Random Occupancy Patterns	85.8
[17] HON4D	92.4
Depth Motion Volumes	95.4
Depth Motion Volumes + Key frames	96.1

The confusion matrix of the average accuracy performed on MSR Gesture3D dataset is listed in Table 2. It can be observed that the proposed approach performs well for most gesture signs. The confusion occurred in some familiar gesture signs. Four signs are correctly the recognized without error, while other four signs (‘*J*’, ‘*Past*’, ‘*Hungary*’, and ‘*Milk*’) have one misclassified each. The confusion occurred in some familiar gesture signs, such as ‘*Pig*’, ‘Z’, ‘*Store*’ and ‘*Green*’. The visual similarity the invisibility of the finger(s) reduced the classification accuracy significantly.

Table 2. The confusion matrix of the experiments

Gesture name	Total sequences	Correctly classified	Recognition rate (%)
Z	28	26	92.86
J	28	27	96.43
Where	28	28	100
Store	28	26	92.86
Pig	25*	23*	92.0*
	* three blank sequences are excluded		
Past	28	27	96.43
Hungary	28	27	96.43
Green	28	25	89.29
Finish	28	28	100
Blue	28	28	100
Bathroom	28	28	100
Milk	28	27	96.43

5.Conclusion

This paper proposes a dynamic sign language recognition approach through the use of Depth Motion Volumes (DMV). Different from the previous research by stacking the difference of the motion energy of depth images, a novel strategy is applied for forming DMV. The unique shape patterns and properties of the 3D object surfaces expose plentiful distinctive information for characterizing category. To speeds up the recognition system, a key frame selection method is adopted in this paper. The selected key frames of the entire video sequences are stacked in chronological order and then projected into three orthogonal planes. Distinctive features are calculated by Histograms of Oriented Gradients (HOG) and then input to LIBSVM for recognition. The experiments are executed on MSR Gesture3D dataset and the results indicate that the proposed approach is efficient and performs better than the compared approaches.

References

1. Roanna Lun and Wenbing Zhao. A Survey of Applications and Human Motion Recognition with Microsoft Kinect. International Journal of Pattern Recognition and Artificial Intelligence. Volume 29, No. 05, 1555008 (2015)
2. Biplab Ketan Chakraborty, Debajit Sarma, M K Bhuyan, et. al. Review of constraints on vision-based gesture recognition for human-computer interaction. IET Computer Vision. 2018, Volume 12, Issue. 1, pp. 3-15
3. Ming Jin Cheok, Zaid Omar, Mohamed Hisham Jaward. A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics. 2019, Volume 10, Issue 1, pp. 131–153
4. Chenyang Zhang and Yingli Tian. Histogram of 3D Facets: A Depth Descriptor for Human Action and Hand Gesture Recognition. Computer Vision and Image Understanding. Volume 139, October 2015, Pages 29-39
5. Wenwen Yang, Jinxu Tao, Zhongfu Ye. Continuous sign language recognition using level building based on fast hidden Markov model. Pattern Recognition Letters. 78 (2016) 28–35
6. Guillaume Plouffe and Ana-Maria Cretu. Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping. IEEE Transactions on Instrumentation and Measurement. VOL. 65, NO. 2, FEBRUARY 2016, 305-316
7. Greg C. Lee, Fu-Hao Yeh, Yi-Han Hsiao. Kinect-based Taiwanese sign language recognition system. Multimedia Tools and Application. (2016) 75:261–279
8. Pradeep Kumara, Himaanshu Gaubaa, Partha Pratim Roya, et. al. Coupled HMM-based multi-sensor data fusion for sign language recognition. Pattern Recognition Letters. 86 (2017) 1–8
9. Shiliang Huang, Chensi Mao, Jinxu Tao, and Zhongfu Ye. A Novel Chinese Sign Language Recognition Method Based on Keyframe-Centered Clips. IEEE Signal Processing Letters. VOL. 25, NO. 3, 2018, 442-446
10. Xiaodong Yang, Chenyang Zhang, Yingli Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. Proceedings of ACM International Conference on Multimedia, pp. 1057–1060, Nara, Japan (2012)
11. Chen Chen, Kui Liu and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps [J]. Journal of Real-Time Image Processing, (2016) 12: 155-163.
12. Kumud Tripathi, Neha Baranwal and G. C. Nandi. Continuous Indian Sign Language Gesture Recognition and Sentence Formation [J]. Procedia Computer Science, 2015, 54: 523-531.
13. Rajeshree S. Rokade and Dharmpal D. Doye. Spelled sign word recognition using key frame [J]. IET Image Processing, 2015, 9(5): 381-388.
14. Ali Asghar Zare and Seyed Hamid Zahiri. Recognition of a real-time signer-independent static Farsi sign language based on Fourier coefficients amplitude [J], International Journal of Machine Learning and Cybernetics. (2018) 9:727–741
15. Alexey Kurakin, Zhengyou Zhang, Zicheng Liu. A Real-Time System for Dynamic Hand Gesture Recognition with a Depth Sensor. 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 27-31 Aug. 2012, Bucharest, Romania, pp. 1975-1979.
16. Jiang Wang, Zicheng Liu, Jan Chorowski, et al. Robust 3D Action Recognition with Random Occupancy Patterns. 12th European Conference on Computer Vision (ECCV), Florence, Italy, October 7-13, 2012, pp. 872-885.
17. Omar Oreifej, Zicheng Liu, HON4D: histogram of oriented 4D normal for activity recognition from depth sequences. 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA (June 23-28) 2013, pp. 716-723.
18. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed on 9 January 2020